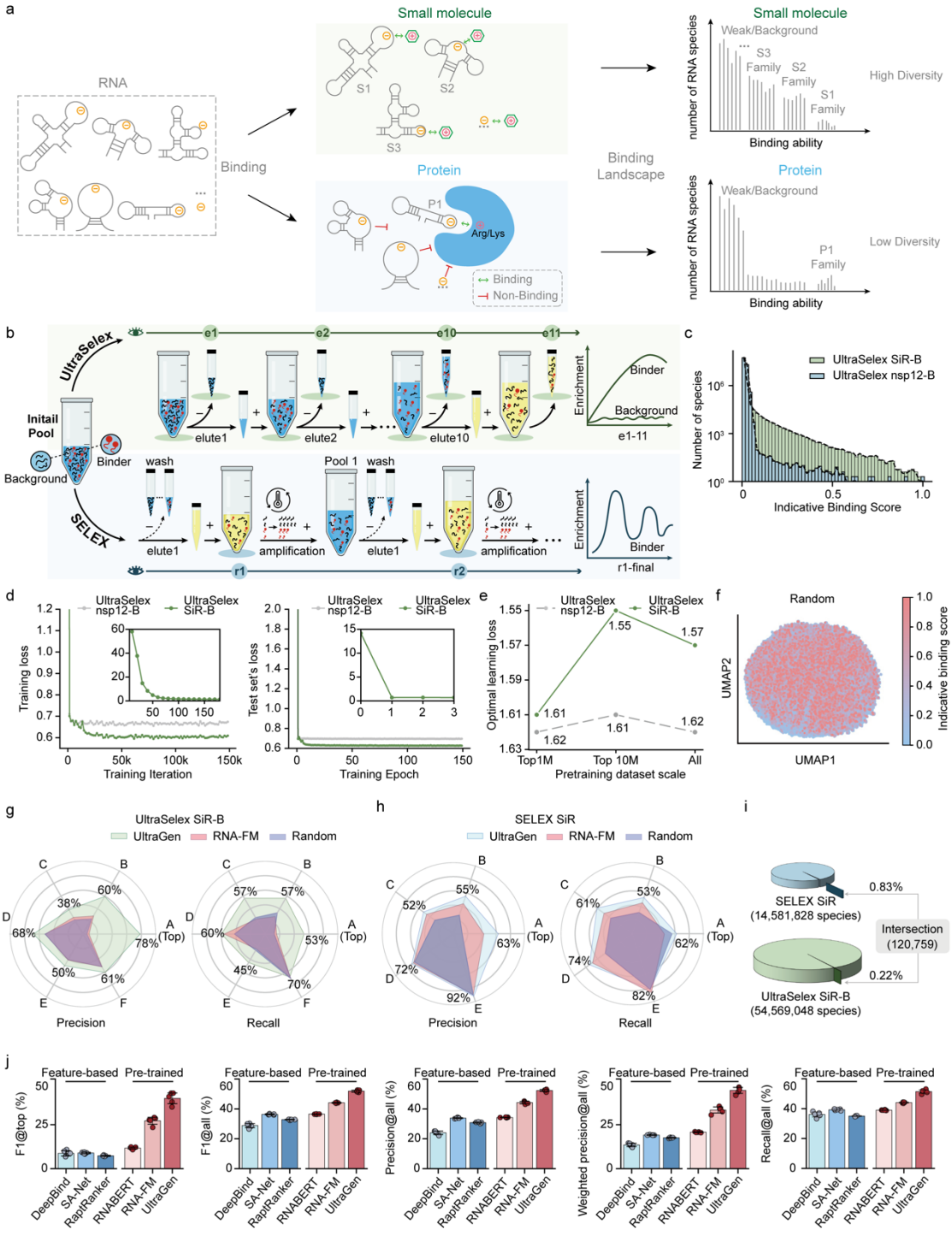
Supplementary Materials

Decoding the RNA binding systems by UltraGen

Hui Wang[1,6], Zhaoming Chen[1,6], Wenjun Lin[1], Yuan Jiang[2], Jingye Zhang[2], Wenhao Huang[1], Yonggui Fu[4], Hongwang Xiao[1,5], David Kuster[3], Andres Jäschke[2], Qiwei Ye[1]*, Yaqing Zhang[1,2]*

[6] These authors contributed equally to this work.

*Correspondence: qwye@baai.ac.cn and yaqing.zhang@uni-heidelberg.de.



Extended Data Fig. 1: Pre-training UltraGen on various RNA binding systems and comparing model performance.

a, Schematic illustration of RNA interactions with small ligands and proteins based on their electrostatic properties, determined by the positively charge. The target-derived geometric constraints result in differences in the diversity of the binding landscape between interactions with small molecules and proteins.

b, Differential enrichment pathway of RNA-target binding between SELEX and UltraSelex[1] techniques. SELEX discontinuously enriches RNA binders, with certain losses occurring during

washing steps across multiple rounds. This iterated process introduces intermediate enzymatic amplification biases, such as T7 RNA polymerase and reverse transcriptase, which leverage RNA abundance regardless of binding strength. In contrast, UltraSelex exhibits a continuous monotonic distribution of enriching RNA binders in a systematic gradient trend within a single round.

c, Distribution of RNA species based on the normalized UltraSelex SGREELI *auc* values, here referred to as "Indicative binding score", compared between UltraSelex silicon rhodamine molecule (SiR)-B and UltraSelex non-structure protein 12 (nsp12)-B RNA dataset.

d, UltraGen's average learning loss on the training (left) and test (right) set of UltraSelex SiR-B and UltraSelex nsp12-B RNA dataset throughout the learning process. The inset line plots depict the learning loss in the initial training phase. The held-out test set comprised 2% of the total sequences, ensuring less than 90% identity to the training set as determined by CD-HIT.

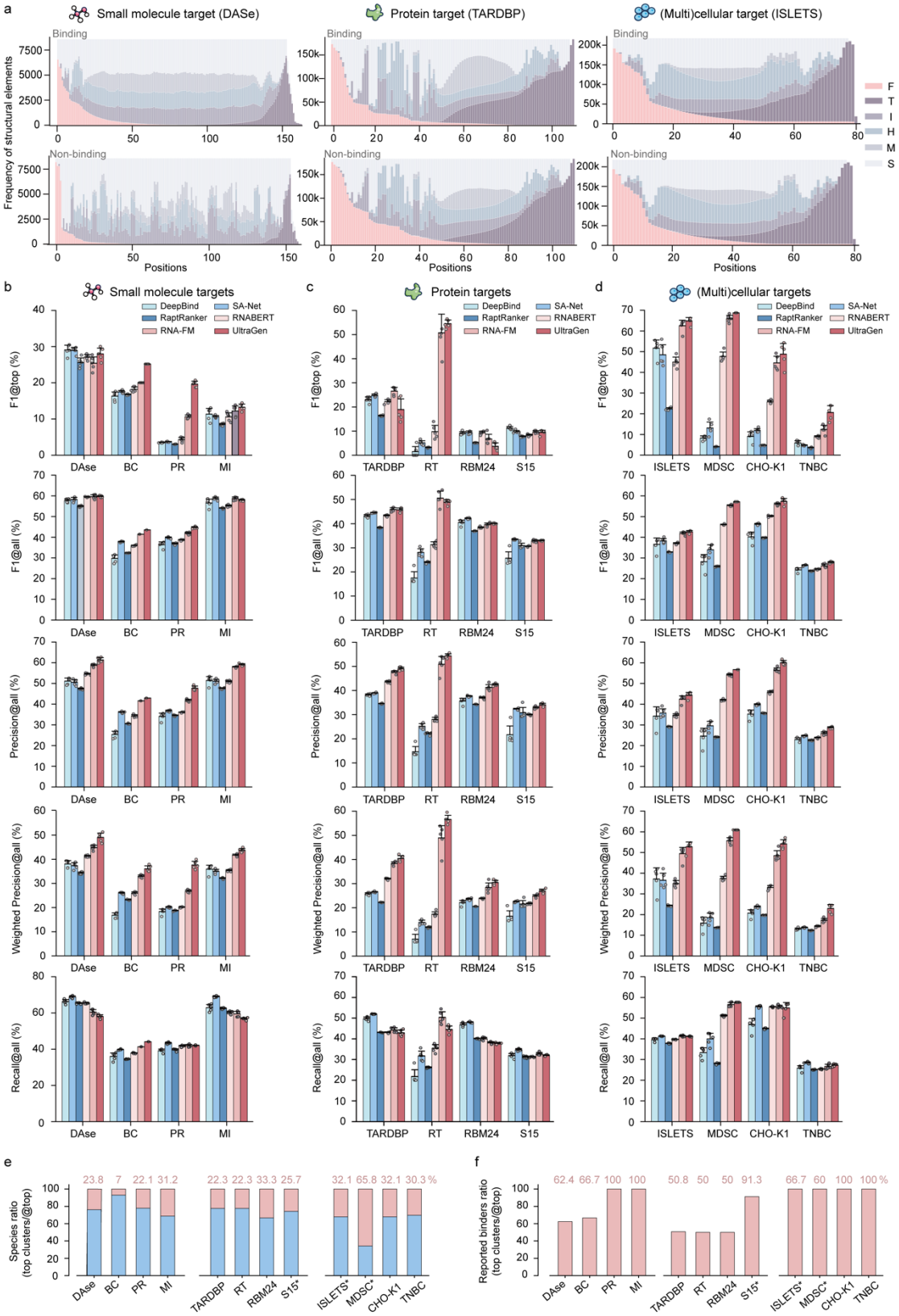
e, Comparing the exponential learning loss on the held-out test sets for different sizes of the training set. The UltraSelex SiR-B and UltraSelex nsp12-B dataset were analyzed and sorted by SGREELI *auc* value in descending order for corresponding binning. The held-out test set was constructed as same in panel d.

f, UMAP projections of top one million RNA binders from UltraSelex SiR-B by randomly initialized UltraGen, colored by UltraSelex SGREELI *auc* values ("Indicative Binding Score").

g&h, Model performance metrics in ranking the held-out test sets of UltraSelex SiR-B system (**g**) and SELEX SiR system (**h**), respectively. Pre-trained UltraGen ("UltraGen", depicted in red or blue) was compared against the RNA-FM (depicted in salmon) and randomly initialized UltraGen ("Random", depicted in light purple). The UltraSelex SiR-B RNA set, ordered alphabetically from A (top 0.01%), B (top 0.01-0.05%), C (top 0.05-0.1%), D (top 0.1-0.5%), E (top 0.5-1%) to F(top 1-100%), was based on the UltraSelex SGREELI *auc* values. The SELEX SiR RNA set, ordered from A (top 1%), B (top 1-5%), C (top 5-12%), D (12-100%), E (Background, no detection on the final round), was based on the sequence abundance in the final round (the 14^{th}). For classification details, see Supplementary Table 1.

i, Marginal intersection between UltraSelex SiR-B and SELEX SiR RNA datasets. All overlapped identical species were excluded from the SELEX SiR dataset for model evaluation in Fig. 2c, irrespective of their binding labels. The detailed counts of RNA species, including those from the downsampling strategy for data balance in each binding category, are provided in Supplementary Table 1.

j, Model ranking comparison for top binders ("@top", identical to that in Fig. 1c) and all binders ("Weighted") between UltraGen and other deep learning models on the held-out test sets. The test sets consist of 30% of the total sequences, containing multiple categories (Supplementary Table 1).



Extended Data Fig. 2: Comprehensive ranking of RNA binders in SELEX systems by UltraGen and other deep learning models.

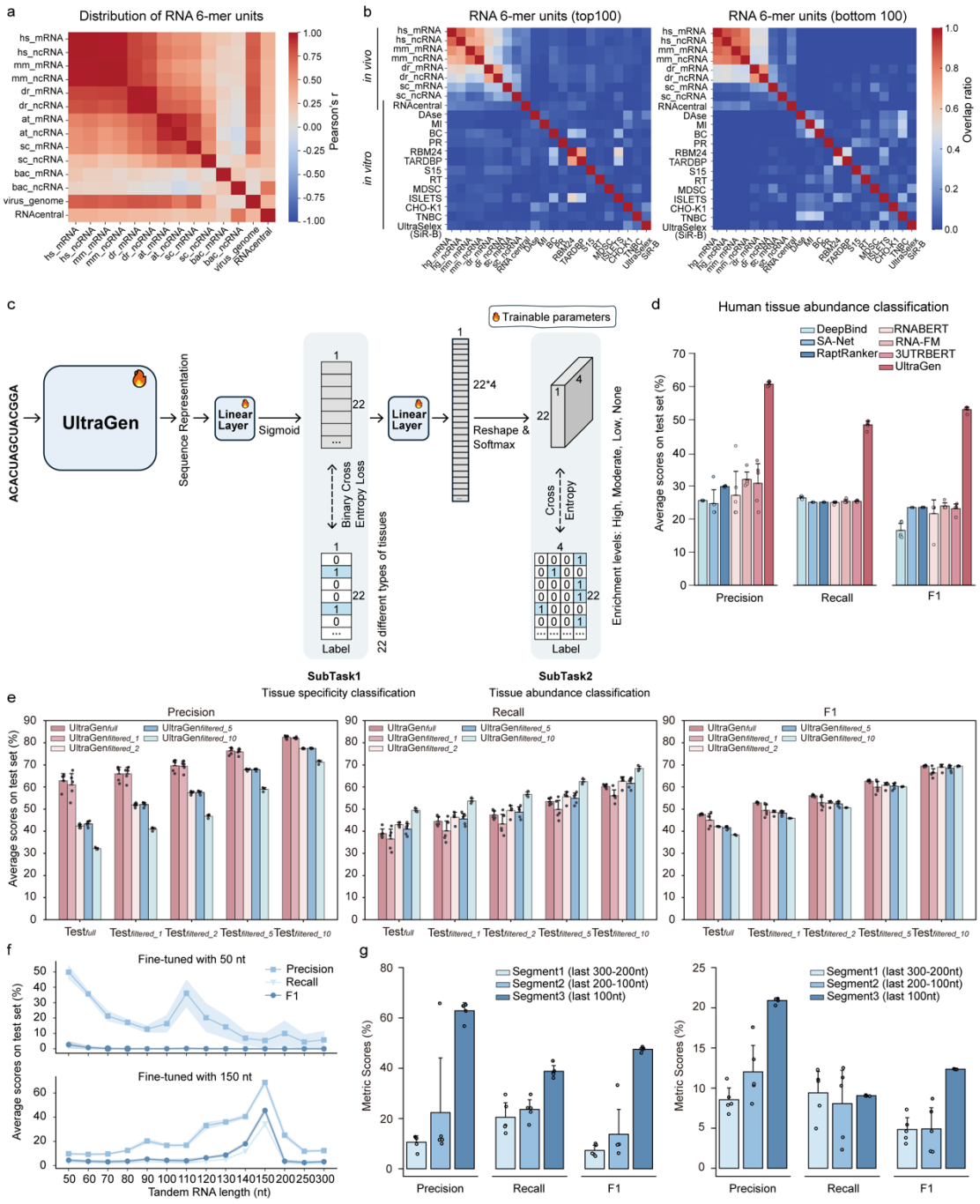
a, Distribution structural discrepancy of RNA species from the SELEX systems in Fig.2a, distinguishing enriched ("Binding") and not enriched ("Non-binding") species. Each nucleotide

of RNA species was labeled using the LinearPartition[2] approach with one of following structural properties: dangling start (F), dangling end (T), internal loop (I), hairpin loop (H), multi-branched loop (M), and stem (S).

b&c&d, Ranking performance on RNA binding system to small molecule (**b**), protein (**c**), and (multi)cellular targets (**d**) between UltraGen and other deep learning models. The evaluated model metrics includes *F1@top*, *F1@all*, *Precision@all*, and *Weighted Precision@all*. The '@top' metrics focus on the top binders, while '@all' metrics address all RNA species. The weighting approach was based on categories with hierarchical binding potential. Full names of targets are in Supplementary Table 2. Error bars represent mean+standard deviation. Model replicates were initiated with independent random seeds, with n=5.

e, The ratio of the RNA species from the top clusters of UltraGen predictions to that in the top category, colored in red. The top five clusters (denoted without "*") and the top twelve clusters (denoted with "*") based on their total HTS counts were summarized.

f, The ratio of known RNA binders in the top clusters of UltraGen predictions (panel d) to that in the top category. The known binders in DAse, MI, TARDBP, RBM24 were inferred from the corresponding reported core binding patterns[3, 4], while the S15 binders[5] were inferred from similar sequences with an edit distance no more than four bases. The known binders of the remaining dataset were directly experimental verified[6-12].



Extended Data Fig. 3: Fine-tuning UltraGen with somatic 3'-UTRs from human tissues.

a, Evolutionarily conserved distribution of RNA hexamer units across species. mRNA and ncRNA from *Homo sapiens* ("hs"), *Mus musculus* ("mm"), *Danio rerio* ("dr"), *Saccharomyces cerevisiae* ("sc"), *Arabidopsis thaliana* ("at"), 5508 random selected bacterial species ("bac"), RNAcentral source ("RNAcentral"), and RNA viruses ("virus_genome") were compared, with colors representing Pearson correlations coefficient ("Pearson's r", two-sided).

b, Disparity in overlap ratio among the top 100 enriched ("top 100", left) and bottom 100 enriched ("bottom 100", right) RNA hexamer units across species. Species abbreviations are consistent with panel a, while SELEX and UltraSelex dataset abbreviations are identical to those in Fig. 2b.

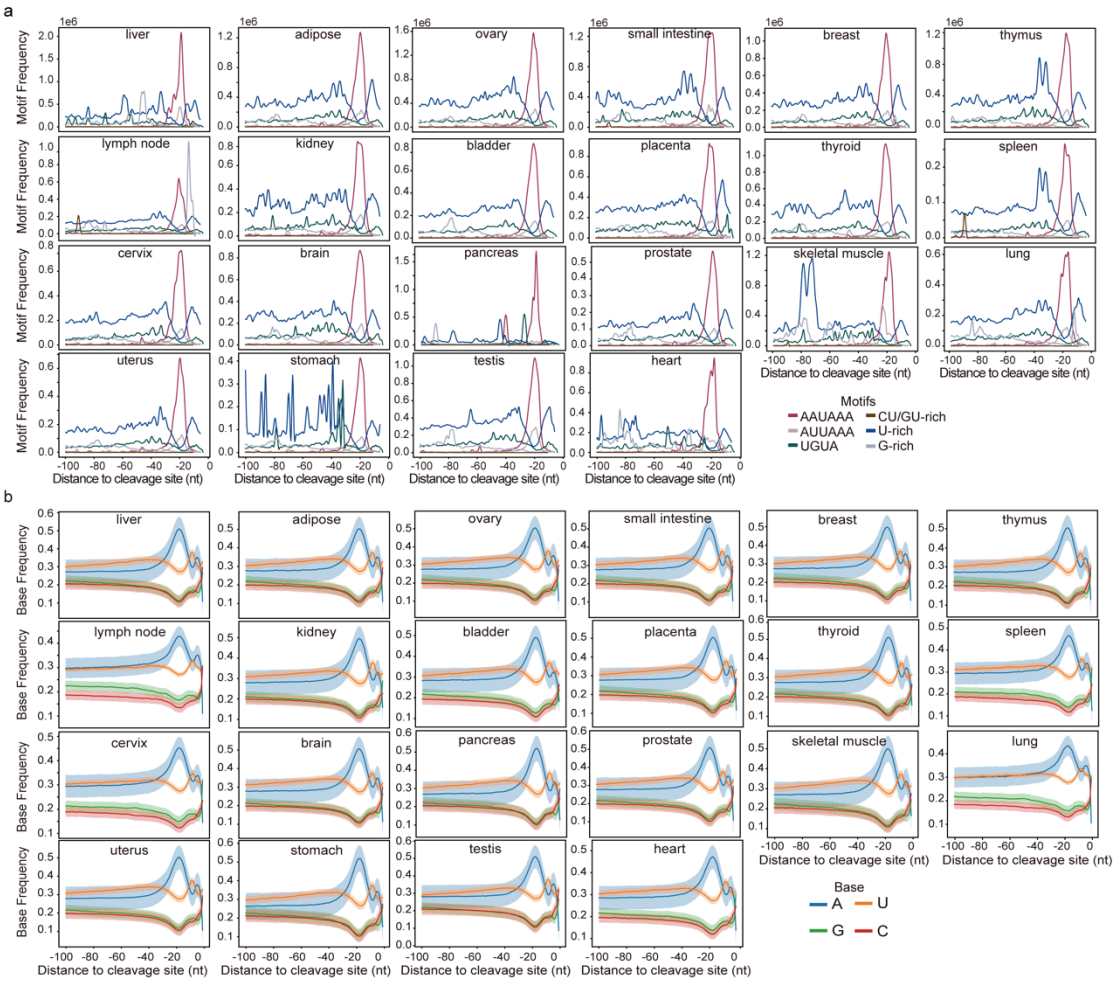
c, A multi-task joint framework of UltraGen integrating 22 human tissue types and four abundance levels. Briefly, during the tissue specificity calculation, each RNA species presents or absence in the specified tissue was predicted with a probability in the range between 0 and 1. RNA species correctly predicted (probability >0.5 for presence, and ≤ 0.5 for absence) were summarized using precision, recall, and F1 score metrics across 22 tissues. RNA species were further classified into four levels: high (counts ≥ 100), intermediate (100 > counts ≥10), low (10 > counts ≥1), and non-existent. The dataset split was performed randomly on the sequences, independent of chromosomes or genes.

d, Identical model performance comparison to Fig. 3a, but focusing on the classification of 3'-UTR abundance levels. Error bars represent mean ± sd. Model replicates were initiated with independent random seeds, n=5.

e, Model performance of UltraGen variants fine-tuned with RNA species from different abundance levels. UltraGen fine-tuned variants with suffixes 'filtered_1', 'filtered_2', 'filtered_5', and 'filtered_10' denote models fine-tuned on datasets excluding RNA species from the training set with maximum abundances across 22 tissues of below 1, 2, 5, and 10 counts, respectively. The test set was also subjected to the same exclusion criteria, removing RNA species that below the defined abundance thresholds. Error bars represent mean + sd. The model replicates is n=5.

f, Length specificity of input RNA for model prediction of tissue classification. The analysis is identical to Fig. 3c, employing tandem 3'-end sequences (50 - 300 nt) with UltraGen model fine-tuned with 50 nt (top panel) and 150 nt (bottom panel) sequences. Error bands indicate sd, with n=5 model replicates.

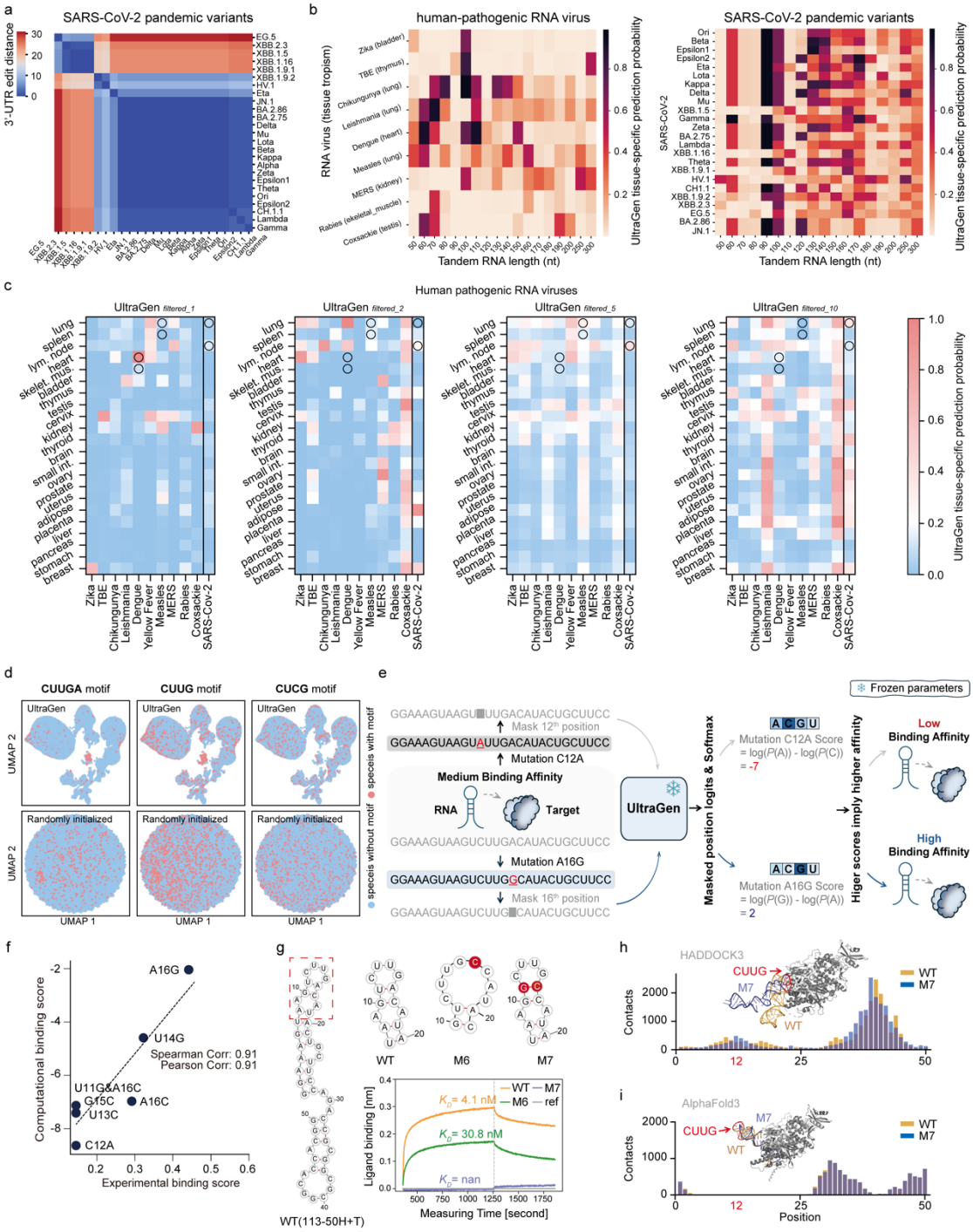
g, Region specificity of input RNA for model prediction of tissue classification based on transcriptome-wide random splitting (left) and chromosome-based split (right). The chromosome-based split follows the same strategy as panel g. The analyzed regions include the last 100 nt, 200-100 nt and 300-200 nt relative to the RNA 3'-end. Error bands indicate sd, with n=5 model replicates.



Extended Data Fig. 4: Characterization of 3'-UTRs from human tissues.

a, Distribution of polyadenylation *cis*-regulatory elements from the upstream region of the cleavage site on each individual 22 human tissues. This analysis is identical to Fig. 3e. RNA species number: liver (219477), adipose (312772), cervix (496468), brain (342821), ovary (365385), small_intestine (378589), pancreas (155310), prostate (330919), breast (282017), thymus (217927), skeletal_muscle (223292), lung (620713), lymph_node (474116), kidney (416171), uterus (381253), stomach (151692), bladder (420678), placenta (320079), testis (628313), heart (346492), thyroid (343288), and spleen (313778).

b, Nucleotide frequency profiles for tandem 3'-end regions in tissue-specific sequences from each individual tissue. RNA species number is identical to panel a. Error band indicates sd.



Extended Data Fig. 5: UltraGen delineates SARS-CoV-2 replicase binding with single-base resolution.

a, Evolutionarily conserved 3'-UTR of SARS-CoV-2 variants over time. The last 100 nucleotides of each 3'-UTR were compared with edit distance. At the genomic level, a mutation rate in spike (23403A>G, 6294/10022 variants) versus 3'-UTR (29870C>A, 115/10022 variants) was reported[13].

b, Length specificity of input tandem RNA for predicting virus tropism. The analysis is identical to Fig. 3c, employing tandem 3'-end sequence from ten endogenous RNA viruses (left) and SARS-CoV-2 variants (right).

c, Attenuated prediction capability of tissue-specific hallmarks in human-pathogenic RNA viruses' 3'-UTRs by UltraGen variants fine-tuned with RNA species of different abundance levels. RNA virus abbreviations are consistent with Fig. 4a,b. Black circles indicate the top two RNA viruses' tropism - dengue, measles, and SARS-CoV-2 - correctly predicted by UltraGen fine-tuned with the full dataset (Supplementary Table 6). UltraGen fine-tuned variants with suffixes 'filtered_1', 'filtered_2', 'filtered_5', and 'filtered_10' denote models fine-tuned on the training dataset excluding RNA species with maximum abundances across 22 tissues of below 1, 2, 5, and 10 counts, respectively.

d, UMAP projections of the top one million RNA binder species from UltraSelex Nsp-B by UltraGen, colored according to the presence or absence of the denoted CUUGA, CUUG, and CUCG motifs.

e, UltraGen's framework for zero-shot inference on SARS-CoV-2 replicase binding. Initially, individual bases are masked. Subsequently, the framework computes the probabilities of predicting bases at the masked positions within the RNA context. This computation relies on the log ratio of the mutant base probability to the wild type base probability."

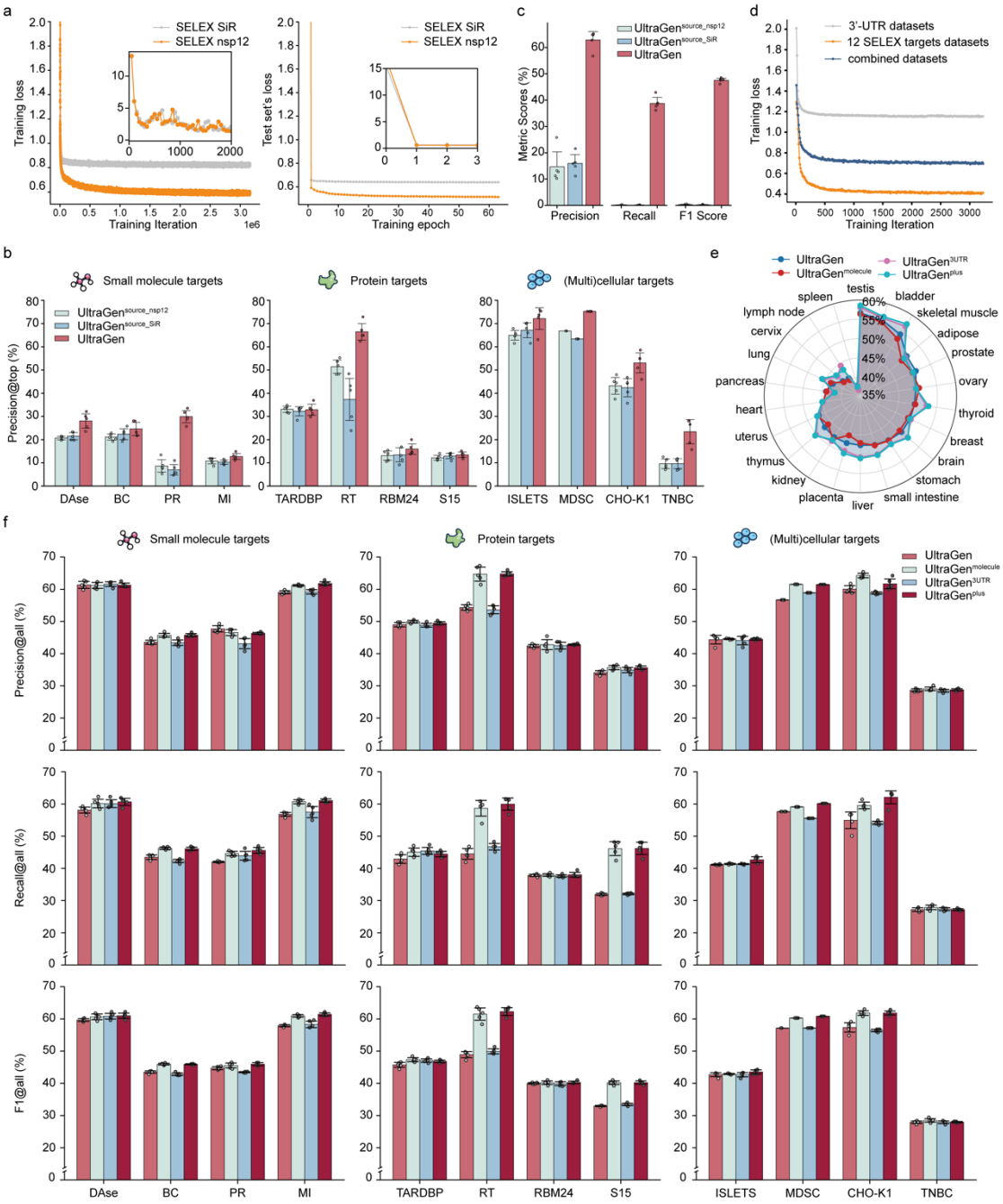
f, Correlation between UltraGen predicted binding score and experimental binding affinity for 'UCUUGA' motif mutation. The dashed line indicates the linear correlation trend. Two-sided.

g, Sequence of wild type (WT, 113-50H+T) RNA and mutants (M6, M7, with mutated bases denoted in red) for studying SARS-CoV-2 replicase binding features. RNA secondary structures were distinguished within the region outlined by a red dashed frame. The reported sequences and K_D values of M1-5 mutants[1] are given in Supplementary Table 8. Line plot shows the bio-layer interferometry measurement of RNA interaction with nsp12 protein. The association phase (left) and dissociates phases (right) were separated by a grey dashed line within the line plot. The reference control ("ref") indicates the detection signal in the absence of RNA during the measurement.

h, Molecular docking of the interaction between SARS-CoV-2 replicase subunit nsp12 protein (PDB: 7btf, chain A) and the tertiary RNA (FARFAR2 sampling). The simulated binding interface of the protein-RNA interaction results was visualized for WT (orange) and M7 (blue) RNA via PyMOL (top panel). A total of 630 interaction simulations were conducted using HADDOCK3[14], including the defined individual (35) or paired (595) combination of reported RNA-interacting amino acids[15]. Individual RNA base-contacting amino acids within the effective distance (<10 Å) were then summed (bottom panel), using the formula $\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{10-distance_{ij}}{10}$. Here, "*n*" represents the number of docking results, "*m*" stands for the interacted amino acid residues, and "$distance_{ij}$" denotes the minimum distance between the alpha carbon atom of the amino acid residue and any heavy atom within the RNA residue. The CUUG structural binding core was colored in red and additionally labeled with its RNA starting position 12. Similar simulation of RNA-protein contacts between WT (binding to nsp12) and M7 (not binding to nsp12) variant with the replicase was carried out.

i, AlphaFold3 (AF3)[16] predictions of the interaction between SARS-CoV-2 replicase subunit nsp12 protein (PDB: 7btf, chain A) and the tertiary RNA (FARFAR2 sampling). Total 100 web server predictions with 20 different random initiation seeds (range from 1000 to 20000, step

size 1000). The contacting amino acids of the predicted results (average pIDDT < 70, AF3 web server interpreted as “Low” confident) were calculated in the same way as in panel d. None of AF3 predictions successfully predicted the experimentally verified CUUG binding core (neither WT nor M7 RNA variant) close to the protein interface, indicating its limitation to identify the correct RNA-protein binding interface.



Extended Data Fig. 6: Model performance with alternative or expanding RNA sources.

a, Pre-training loss on the training (left) and test (right) sets of UltraGen^{source_SiR} and UltraGen^{source_nsp12} model throughout the learning process. Inset line plots depict the initial phase of learning loss. The held-out test set comprised 2% of the total top 10 million sequences, with less than 90% identity to the training set as determined by CD-HIT.

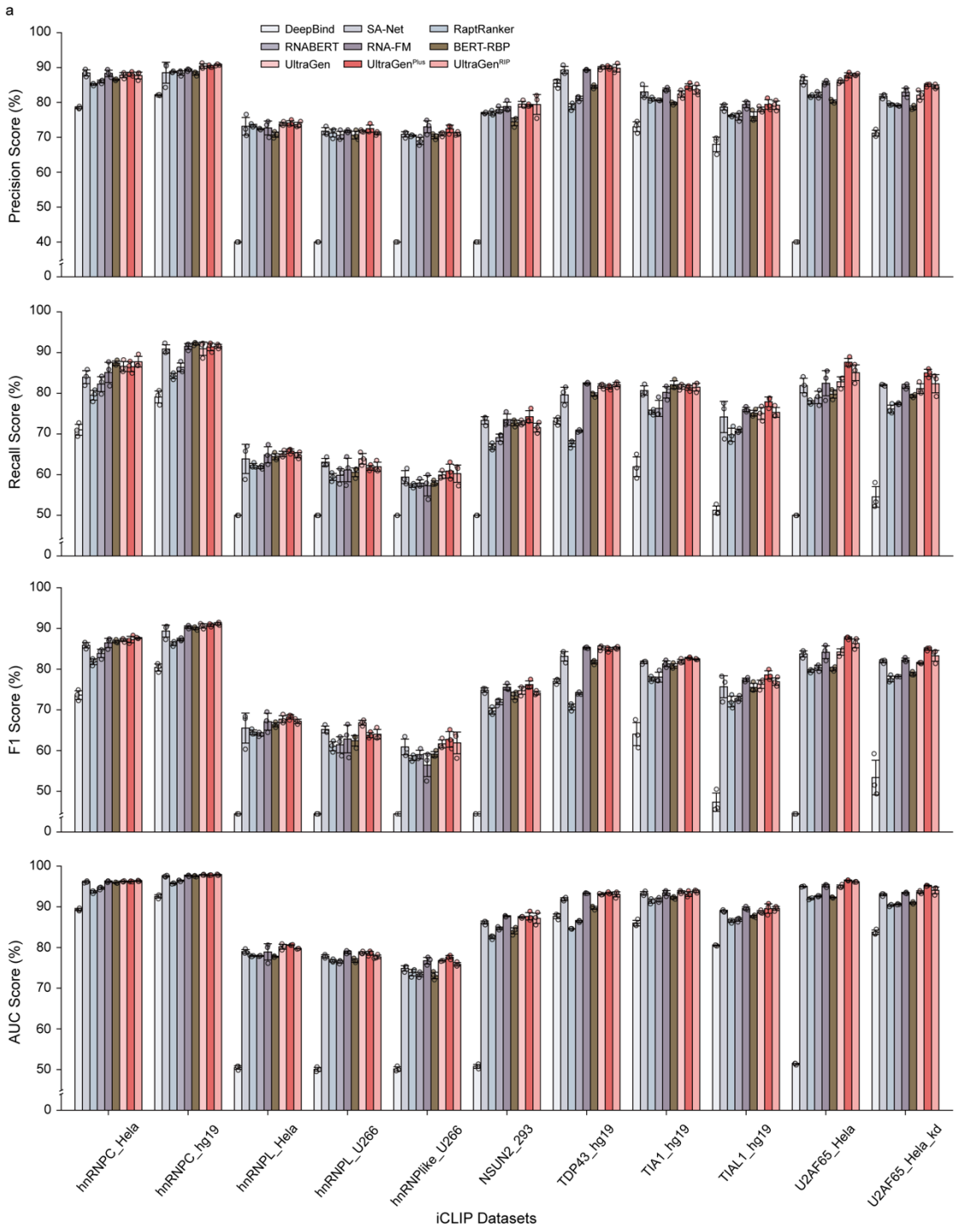
b, Ranking performance of RNA binding system to small molecule, protein, and (multi)cellular targets for UltraGen, UltraGen^{source_SiR}, and UltraGen$^{source_-nsp12}$. The analysis is identical to the Fig. 2b. Error bars represents sd, with n=5 model replicates.

c, Comparison of multi-classification performance for individual 3'-UTRs across 22 human tissues by UltraGen, UltraGen^{source_SiR}, and UltraGen^{source_nsp12} moel. The tissue specificity analysis is identical to Fig. 3a. Error bars represents sd, with n=5 model replicates.

d, Continued pre-training loss for UltraGen variants: UltraGen^{molecules} (the twelve SELEX training dataset), UltraGen^{3UTR} (the preliminary 3'-UTR training dataset), and UltraGen^{plus} (combined SELEX and 3'-UTR training dataset). Each variant underwent an additional round of pre-training to minimize overfitting while retaining core memory from the base UltraGen model.

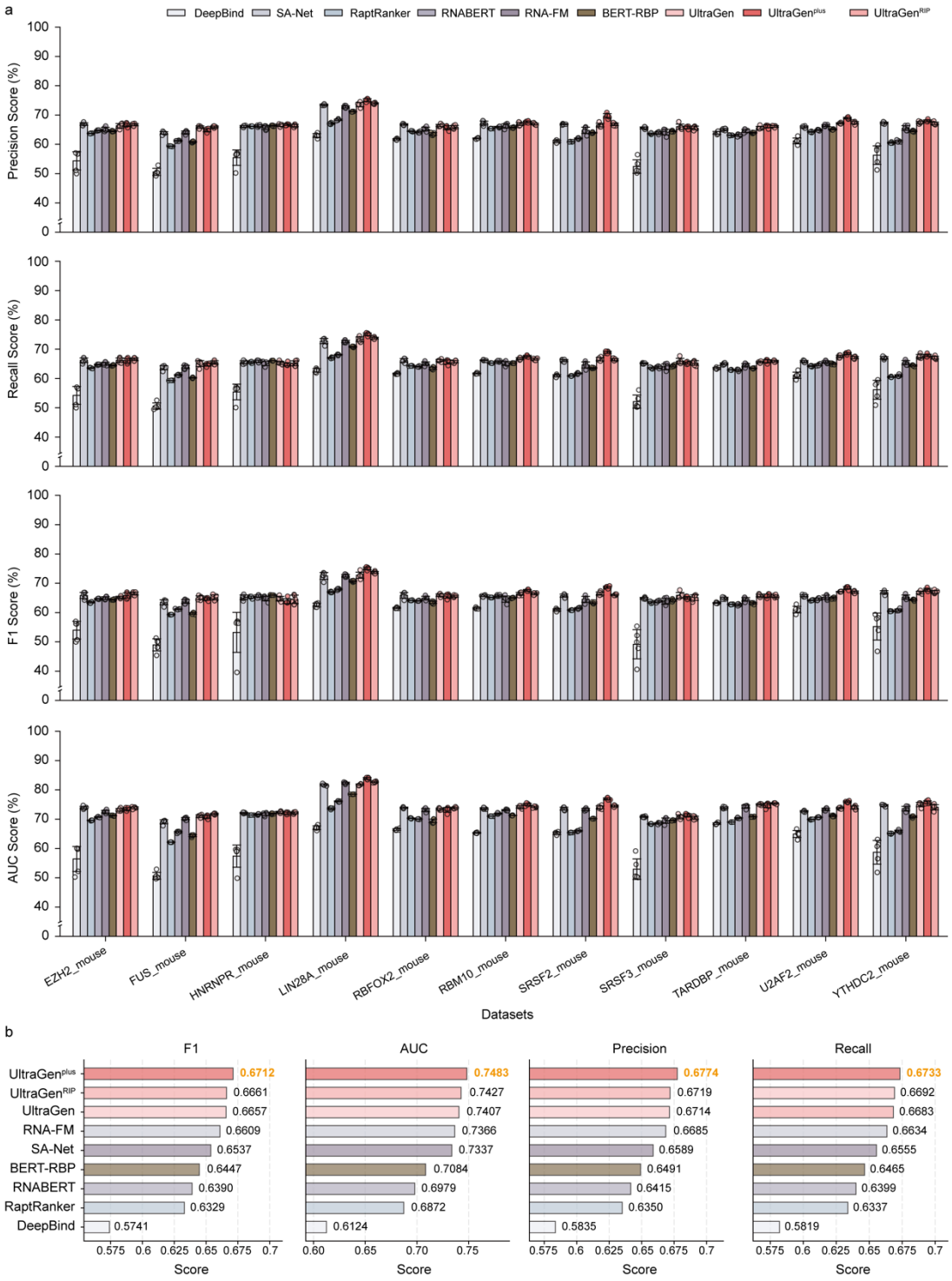
e, Multi-classification F1 score for tissue specificity across individual human tissue by UltraGen, UltraGen^{molecules}, UltraGen^{3UTR}, and UltraGen^{plus} model.

f, Ranking performance on RNA binding system to small molecule, protein, and (multi)cellular targets for UltraGen, UltraGen^{molecules}, UltraGen^{3UTR}, and UltraGen^{plus} model. The evaluated model metrics includes *Precision@all* (top panel), *Recall@all* (middle panel)*, and F1@all* (bottom panel). The analysis is identical to Extended Data Fig. 2a-c. Error bars represent sd, with n=5 model replicates.



Extended Data Fig. 7: Model performance in predicting RNA-protein interaction on the human iCLIP datasets.

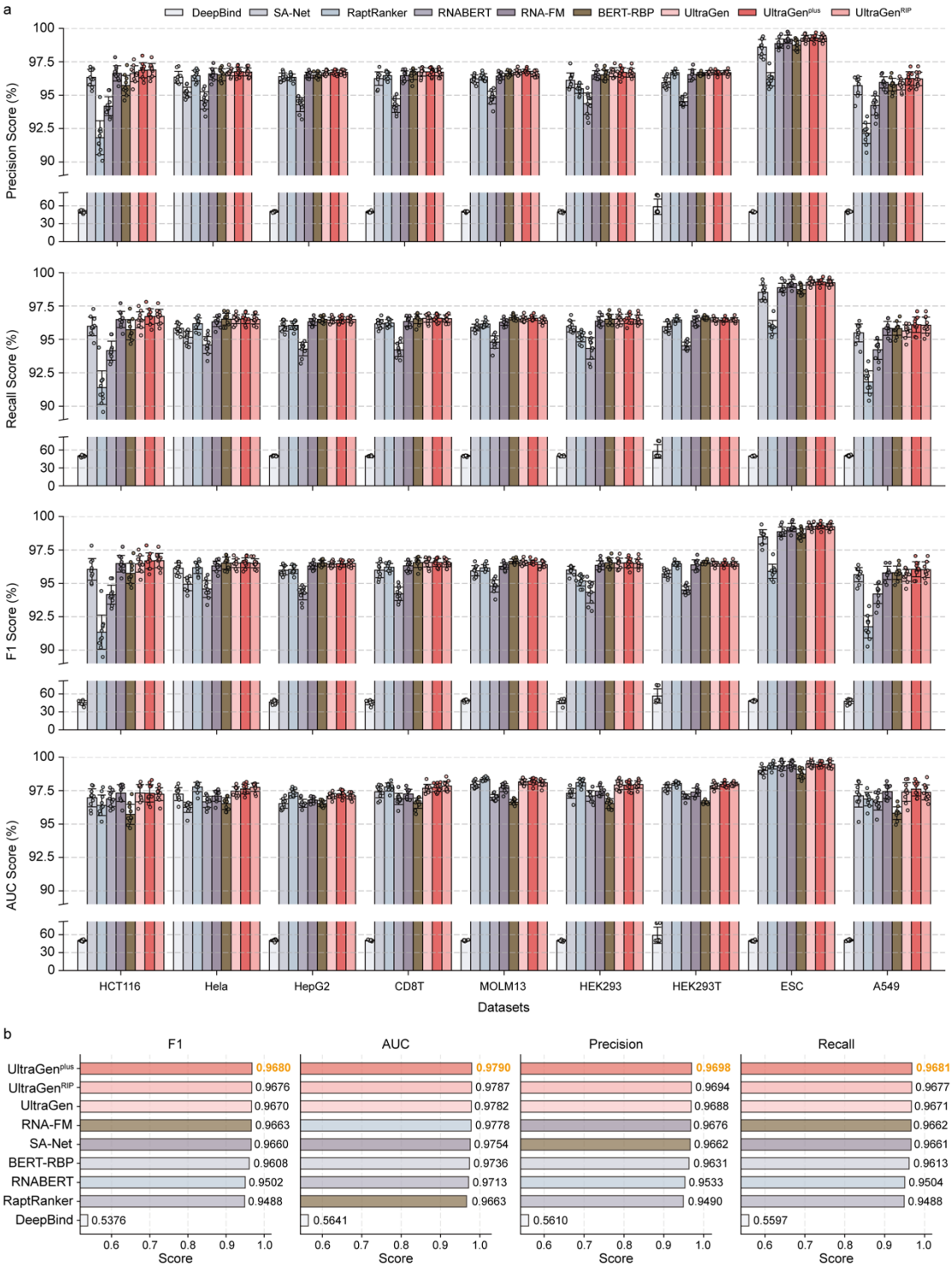
a, Classification performance metrics of UltraGen and other models on the *in vivo* human RNA-protein interaction iCLIP datasets. The eleven iCLIP datasets were curated and individually split into training and testing sets from the iONMF[17] collections. Nine protein targets, including hnRNPC[18], U2AF2[18], hnRNPL[19], hnRNPL-like[20], Nsun2[21], TDP-43[22], TIA1[23], and TIAL1[23] from various cell lines, were examined. The bar height represents the mean performance across the three-fold (n=3) cross-validation, with the standard deviation (sd) indicated by error bar.



Extended Data Fig. 8: Model performance in predicting RNA-protein interaction on the mouse CLIP datasets.

a, Classification performance metrics of UltraGen and other models on the *in vivo* mouse RNA-protein interaction CLIP datasets. The positive sequences of eleven mouse protein targets, including EZH2, FUS, HNRNPR, LIN28A, RBFOX2, RBM10, SRSF2, SRSF3, TARDBP, and YTHDC2, were curated from the CLIPdb[24], while the negative sequences were randomly sampled from transcriptome without overlap with positive sequence. The bar height represents the mean performance (n=5), with the standard deviation (sd) indicated by error bar.

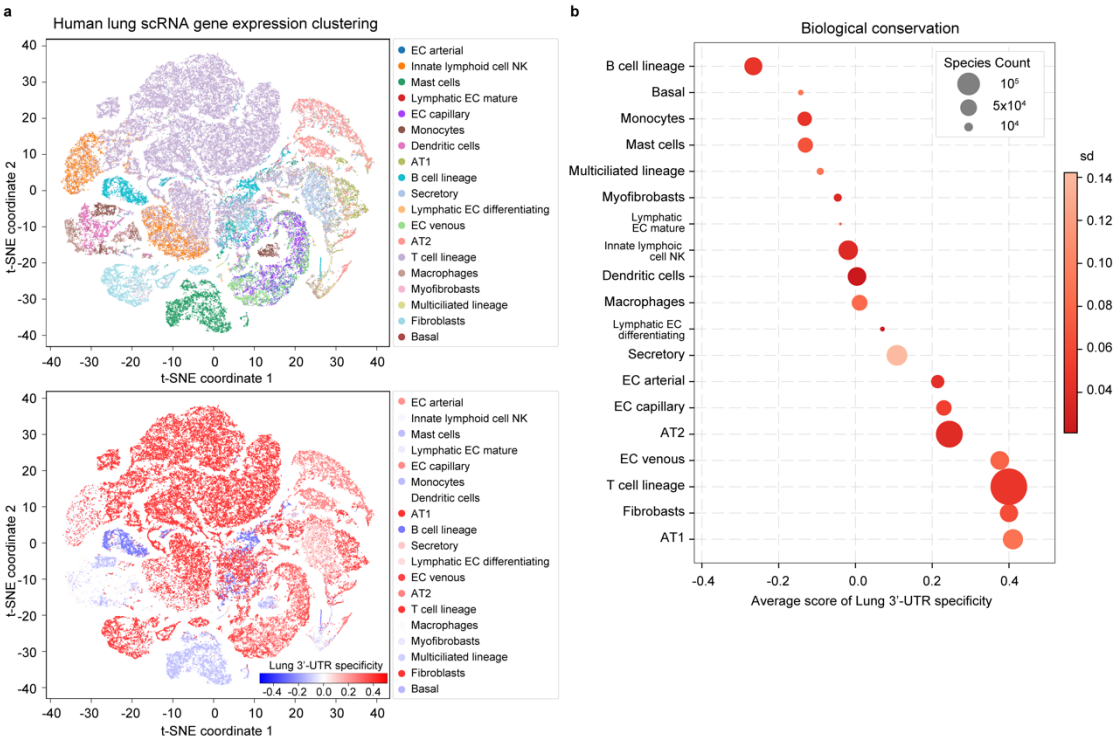
b, Comparison of mouse RNA-protein interactions binary classification from the CLIP datasets between UltraGen variants and other models. The average performance metrics across the eleven CLIP datasets from panel a are presented.



Extended Data Fig. 9: Model performance in predicting human m6A methylation.

a, Classification performance metrics of UltraGen and other models on the *in vivo* RNA m6A methylation datasets. A total of non-redundant 79,021 m6A modification site (filtered from m6A-altas 131,703 raw signals[25]) and 849,005 non-m6A sites, along with their flanking 20-nt upstream and 20-nt downstream regions, from nine cell lines were obtained from a previous study[26]. The bar height represents the mean performance (n=5), with the standard deviation (sd) indicated by error bar.

b, Comparison of RNA m6A methylation binary classification from the CLIP datasets between UltraGen variants and other models. The average performance metrics across the nine datasets from panel a are presented.



Extended Data Fig. 10: Mapping of single-cell 3' readout RNA sequencing of human lung adenocarcinoma by UltraGen.

a, t-SNE projections of human lung adenocarcinoma (LUAD) cells (top panel, expression level based) with annotated lung 3'-UTR specificity (bottom panel, tandem RNA based). Five single cell 3' readout RNA sequencing dataset of primary lung tissues were analyzed. The tissue specificity of RNA species, based on their tandem 100 nt sequences extracted from their mapped loci, was predicted using the UltraGen model. Clustered cells were annotated with the average logistic regression coefficient of five biological samples, with high predicted values indicating lung tissue specificity (the highest classification to lung or probability (range 0-1) above 0.9). Abbreviation, EC (Endothelial cells) and AT (Alveolar type cells).

b, Conservation of lung specific 3'-UTR in alveolar type cells associated with LUAD. Alveolar type II[27] and I[28] cells are highly involved in LUAD progression. The bubble size denotes the number of RNA species for the analysis, and the variance between the five biological replicates is color-coded in red.

Supplementary Table 1. Model ranking criteria for SiR-binding RNA systems.

Dataset	Metrics	Category A	Category B	Category C	Category D	Category E	Category F
UltraSelex SiR-B (whole)	Top Rank*	0.01%	0.01-0.05%	0.05-0.1%	0.1-0.5%	0.5-1%	1-100%
	RNA Species	5,456	21,828	27,285	218,276	272,845	54,023,358
	Random Sample	5,456	21,828	27,285	100,000	100,000	100,000
	Train/Val/Test	3,295/589/1,572	13,101/2,106/6,621	16,422/2,727/8,136	60,045/9,907/3,0048	59,831/10,175/2,9994	60,047/9,953/30,000
SELEX SiR (whole)	Top Rank**	1%	1-5%	5-12%	12%-100%	Background	–
	Enrichment	≥47	[7, 47)	[3, 7)	[1, 3)	0	–
	RNA Species	7,793	32,290	52,699	688,150	13,800,896	–
	Random Sample	7,793	32,290	52,699	100,000	100,000	–
	Train/Val/Test	4,722/750/2,321	19,320/3,332/9,638	31,715/5,286/15,698	59,829/9,936/30,235	60,083/9,974/29,943	–
SELEX SiR (exclusive)	Top Rank**	0.5%	0.5-1%	1-5%	5-16%	16-100%	Background
	Enrichment	≥37	[19, 37)	[5, 19)	[2, 5)	1	0
	RNA Species	3,561	3,842	24,862	82,213	594,381	13,752,210
	Random Sample	3,561	3,842	24,862	82,213	100,000	100,000
	Train/Val/Test	2,134/343/1,084	2,328/382/1,132	14,868/2,461/7,533	49,532/8,232/24,449	59,852/10,112/30,036	59,972/9,918/30,110

* UltraSelex SiR-B RNA species were sorted in a descending order based on SGREELI *auc* values.

**SELEX SiR RNA species were sorted based on their detected enrichment in the final round (the 14^{th}).

Background indicates no detection of these RNA species in the final selection round.

Supplementary Table 2. Summary of twelve benchmark RNA SELEX systems and their high-throughput sequencing libraries.

Target	Name	Dataset	Method	Total rounds	Sequenced rounds	# of species	Length (nt)	PMID
Small molecules	Benzopyrylium-coumarin fluorophores	BC	SELEX	11	1-11	65,299,912	98-106	34309994
	Paromomycin	PR	Capture-SELEX	11	0-11	15,234,700	119-125	30957848
	Maleimide involved in Diels-Alderase	DAse	SELEX	10	1-10	323,414	152-162	24157838
	Mechanistic inhibitor of serine proteases PPACK	MI	SELEX	14	1-14	555,707	228-238	24157838
Proteins	TAR DNA binding protein 43	TARDBP	HTR-SELEX	4	1-4	1,019,953	109	32703884
	Ribosomal protein S15	S15	SELEX	11	11	1,585,543	87	28587636
	RNA-binding motif protein 24	RBM24	HTR-SELEX	4	1-4	749,249	109	32703884
	HIV-1 reverse transcriptase	RT	SELEX	15	15	62,444	115-121	23385524
Multi(cellular) molecules	Triple-negative breast cancer cells	TNBC	Cell-SELEX	14	0, 3, 5, 8, 9, 10, 11, 12, 13, 14	45,922,615	84	32222697
	Chinese hamster ovary K1 cells	CHO-K1	Cell-SELEX	15	0-15	8,432,785	94-99	29982617
	Myeloid-derived suppressor cells	MDSC	Cell-SELEX	11	1,6,10,11	15,297,461	74-78	32554710
	Human islets	ISLETS	Tissue-SELEX	8	1-8	89,694,683	80-84	35383192

Supplementary Table 3. Model ranking criteria for RNA SELEX systems with small-molecule targets.

Dataset	Metrics	Category A	Category B	Category C	Category D	Category E
BC	Top rank*	1%	1-5%	5-15%	15-100%	Background
	Enrichment	≥20	[4, 20)	[2, 4)	1	0
	RNA species	22,034	97,948	200,504	1,829,680	63,149,746
	Random Sample	22,034	97,948	100,000	100,000	100,000
	Train/Val/Test	13,213/2,191/6,630	58,790/9,828/29,330	59,936/10,038/30,026	59,970/9,959/30,071	60,080/9,982/29,938
PR	Top rank*	1%	1-5%	5-10%	10-100%	Background
	Enrichment	≥50	[7, 50)	[3, 7)	[1, 3)	0
	RNA species	1,410	5,714	9,060	124,388	15,094,128
	Random Sample	1,410	5,714	9,060	100,000	100,000
	Train/Val/Test	853/135/422	3,451/619/1,644	5,407/914/2,739	60,017/10,031/29,952	59,982/9,919/30,099
DAse	Top rank*	20%	20-100%	Background	–	–
	Enrichment	≥2	1	0	–	–
	RNA species	1,720	6,863	314,831	–	–
	Random Sample	1,720	6,863	100,000	–	–
	Train/Val/Test	1,020/164/536	4,152/652/2,059	59,977/10,042/29,981	–	–
MI	Top rank*	3%	3-100%	Background	–	–
	Enrichment	≥2	1	0	–	–
	RNA species	924	33,635	521,148	–	–
	Random Sample	924	33,635	100,000	–	–
	Train/Val/Test	542/83/299	20,173/3,317/10,145	60,020/10,056/29,924	–	–

* SELEX RNA species were sorted based on their detected enrichment in the final SELEX round. Background indicates no detection of these RNA species in the final selection round.

Supplementary Table 4. Model ranking criterial for RNA SELEX systems with protein targets.

Dataset	Metrics	Category A	Category B	Category C	Category D	Category E
TARDBP	Top rank*	4%	4-12%	12-100%	Background	_
	Enrichment	≥3	[2, 3)	[1, 2)	0	_
	RNA species	7,985	13,314	162,401	836,253	_
	Random Sample	7,985	13,314	100,000	100,000	_
	Train/Val/Test	4,814/775/2,396	8,026/1,358/3,930	60,021/9,909/30,070	59,918/10,088/29,994	_
S15	Top rank*	1%	1-5%	5-11%	11-100%	_
	Enrichment	≥38	[6, 38)	[3, 6)	[1, 3)	_
	RNA species	6,720	68,956	99,348	1,410,519	_
	Random Sample	6,720	68,956	99,348	100,000	_
	Train/Val/Test	4,059/721/1,940	41,289/6,882/20,785	59,779/9,863/29,706	59,887/10,036/30,077	_
RBM24	Top rank*	1%	1-6%	6-100%	Background	_
	Enrichment	≥3	[2, 3)	[1, 2)	0	_
	RNA species	2,365	8,226	170,684	567,974	_
	Random Sample	2,365	8,226	100,000	100,000	_
	Train/Val/Test	1,420/229/716	4,946/833/2,447	60,034/10,056/29,910	59,954/9,941/30,105	_
RT	Top rank*	1%	1-5%	5-10%	10-31%	31-100%
	Enrichment	≥153	[17, 153)	[6, 17)	[2, 6)	[1, 2)
	RNA species	625	2,496	3,783	13,340	42,200
	Random Sample	625	2,496	3,783	13,340	42,200
	Train/Val/Test	372/63/190	1,480/247/769	2,294/341/1,148	7,989/1,347/4,004	25,331/4,246/12,623

* SELEX RNA species were sorted based on their detected enrichment in the final SELEX round. Background indicates no detection of these RNA species in the final selection round.

Supplementary Table 5. Model ranking criteria for RNA SELEX systems with (multi)cellular targets.

Data set	Metrics	Category A	Category B	Category C	Category D	Category E	Category F	Category G
TNBC	Top rank*	0.1%	0.1-0.5%	0.5-1%	1-5%	5-9%	9-100%	Background
	Enrichment	≥115	[26, 115)	[14, 26)	[5, 14)	[4, 5)	[1, 4)	0
	RNA species	1,711	6,931	9,455	64,549	70,968	1,547,821	44,221,180
	Random Sample	1,711	6,931	9,455	64,549	70,968	100,000	100,000
	Train/Val/Test	1,025/164/522	4,181/703/2047	5,663/965/2,827	38,800/6,547/19,202	42,615/7060/21,293	59,805/9995/30,200	60,079/9,927/29,994
CHO-K1	Top rank*	1%	1-5%	5-12%	12-100%	Background	–	–
	Enrichment	≥32	[4, 32)	[2, 4)	[1, 2)	0	–	–
	RNA species	1,250	5,432	8,561	108,438	8,309,104	–	–
	Random Sample	1,250	5,432	8,561	100,000	100,000	–	–
	Train/Val/Test	753/120/377	3,287/535/1,610	5,116/820/2625	60,059/9,964/29,977	59,930/10,085/29,985	–	–
MDSC	Top rank*	0.1%	0.1-0.5%	0.5-1%	1-5%	5-10%	10-100%	Background
	Enrichment	≥165	[30, 165)	[14, 30)	[3, 14)	[2, 3)	[1, 2)	0
	RNA species	3,642	14,698	19,087	158,808	161,512	3,277,166	11,662,548
	Random Sample	3,642	14,698	19,087	100,000	100,000	100,000	100,000
	Train/Val/Test	2,226/368/1,048	8,751/1,498/4,449	11,465/1,907/5,715	59,948/9,918/30,134	59,939/10,059/30,002	60,094/9,984/29,922	60,033/1,0008/29,959
ISLETS	Top rank*	0.1%	0.1-2%	2-100%	Background	–	–	–
	Enrichment	>=3	[2, 3)	[1, 2)	0	–	–	–
	RNA species	13,599	373,679	17,547,704	71,759,701	–	–	–
	Random Sample	13,599	100,000	100,000	100,000	–	–	–
	Train/Val/Test	8,206/1,319/4,074	60,155/9,864/29,981	59,858/10,048/30,094	59,940/10,129/29,931	–	–	–

* SELEX RNA species were sorted based on their detected enrichment in the final SELEX round. Background indicates no detection of these RNA species in the final selection round.

Supplementary Table 6. Summary of sequence usage for UltraGen and relevant variants fine-tuned with tandem 3'-UTR RNA species from 22 human tissues.

UltraGen Variants	Train species	Eval species	Test species	Total	Percentage
Full*	1,688,633	361,168	844,317	2,894,118	100%
Filtered_1**	1,030,127	218,650	514,684	1,716,472	60.92%
Filtered_2**	757,708	160,587	378,635	1,262,869	44.81%
Filtered_5**	417,312	88,356	208,635	695,787	24.68%
Filtered_10**	230,537	48,708	115,191	384,160	13.63%

* A full collection of tandem 3'-terminal end sequencing data across 22 human tissues obtained from APAdb was segregated into training, validation, and test sets with an approximate 6:1:3 distribution.

** The training, validation, and testing sets underwent progressive filtering using four abundance cutoffs (>1, >2, >5, and >10), where sequences were retained only if their maximum abundance across 22 tissues exceeded the respective threshold.

Supplementary Table 7. Summary of human-pathogenic RNA viruses and their tissue-specific inference.

Name	Abbreviation	Accession ID	PolyA length	3'-UTR*	Tissue top2 prediction	Tissue literature (PMID)
Zika virus	Zika	NC_012532.1	0	100	bladder, lymph node	29494684
Tick-borne encephalitis virus	TBE	NC_001672.1	0	100	thymus, bladder	2389567, 36205381
Chikungunya virus	Chikungunya	NC_004162.2	0	100	lung, lymph node	36016408
Leishmania RNA virus 1 - 1	Leishmania	NC_002063.1	0	36	lung, spleen	31999729
Dengue virus 1	Dengue	NC_001477.1	0	100	heart, skeletal muscle	20032806
Yellow fever virus	Yellow Fever	NC_002031.1	0	100	lymph node, spleen	30134625, 25412185
Measles virus	Measles	NC_001498.1	0	97	lung, spleen	17715217, 10490102
Middle East respiratory syndrome-related coronavirus	MERS	NC_019843.3	12	100	kidney, spleen	31668197, 26203058
Rabies virus	Rabies	NC_001542.1	0	74	skeletal muscle, heart	4282379, 33738007
Human coxsackievirus A2 strain Fleetwood	Coxsackie	NC_038306.1	0	71	testis, thyroid	33102086, 35909527

* The length of sequence from the end of the 3'-UTR that was utilized for model prediction.

299 **Supplementary Table 8. Summary of SARS-CoV-2 variants for tissue-specific inference.**

Name	Lineage	Accession ID	PolyA length	3'-UTR*
Ori	B	NC_045512.2	33	100
Epsilon2	B.1.429	MZ722827.1	33	100
Zeta	P.3	OL981152.1	33	100
Epsilon1	B.1.427	OK245663.1	33	100
Theta	P.2	MZ169917.1	33	100
Eta	B.1.525	MW791296.1	33	100
Lambda	C.37	MW937858.1	33	100
Gamma	P.1	OP855417.1	33	100
Alpha	B.1.1.7	OX590588.1	33	100
Kappa	B.1.617.1	OK246830.1	33	100
Beta	B.1.351	OK246830.1	33	100
Lota	B.1.526	OK246905.1	33	100
Mu	B.1.621	MZ411658.1	33	100
Delta	B.1.617.2	OR129998.1	33	100
Omicron_multiple	BA.2.75	OP998291.1	33	100
Omincron_XBB_1.16_multiple	XBB.1.16	BS007362.1	21	100
Omicron_CH1.1_multiple	CH.1.1	OR180688.1	33	100
Omincron_XBB_1.9.1_multiple	XBB.1.9.1	BS007360.1	21	100
Omincron_XBB_1.5_multiple	XBB.1.5	BS007567.1	21	100
Omincron_XBB_1.9.2_multiple	XBB.1.9.2	BS007596.1	30	100
Omincron_XBB_2.3_multiple	XBB.2.3	BS007565.1	21	100
BA.2.86_multiple	BA.2.86	OR558991.1	33	100
Omicron_EG.5_multiple	EG.5	BS007694.1	24	100
HV.1	HV.1	OR815854.1	29	100
JN.1	JN.1	OR815892.1	33	100

300 * The length of the sequence from the end of 3'-UTR that was utilized for model prediction.

301

Supplementary Table 9. Consistence between experimental and computational methods for nsp12 RNA binding.

113-50H+L	Mutation	K_D (nM)	Experimental binding*	Computational binding**	RNA Sequence
WT	_	4.1	_	_	GGAAAGUAAGUCUUGACAUACUGCUUCCAGACCGCGGCGCGGCAC CACGG
M1	C12A	no binding (>1000)	0.1448	-8.6362	GGAAAGUAAGU*A*UUGACAUACUGCUUCCAGACCGCGGCGCGGCAC CACGG
M2	U13C	no binding (>1000)	0.1448	-7.4149	GGAAAGUAAGUC*C*UGACAUACUGCUUCCAGACCGCGGCGCGGCAC CACGG
M3	U14C	22.2	0.3226	-4.5929	GGAAAGUAAGUCU*C*GACAUACUGCUUCCAGACCGCGGCGCGGCAC CACGG
M4	G15C	no binding (>1000)	0.1448	-7.4056	GGAAAGUAAGUCUU*C*ACAUACUGCUUCCAGACCGCGGCGCGGCAC CACGG
M5	A16G	9.6	0.4421	-2.0393	GGAAAGUAAGUCUUG*G*CAUACUGCUUCCAGACCGCGGCGCGGCAC CACGG
M6	A16C	30.8	0.2918	-6.9735	GGAAAGUAAGUCUUG*C*CAUACUGCUUCCAGACCGCGGCGCGGCAC CACGG
M7	U11G&A16C	no binding (>1000)	0.1448	-7.1336	GGAAAGUAAG*G*CUUG*C*CAUACUGCUUCCAGACCGCGGCGCGGCAC CACGG

* Experimental binding scores were calculated as $1/\ln(K_D)$.

** Computational binding scores, with higher values indicating greater binding potential, were predicted by UltraGen.

Supplementary Table 10. Summary of human iCLIP sequencing datasets

Dataset*	Target Protein	Tissue	Protocol	PMID	Data file name
hnRNPC_Hela	hnRNPC	HeLa	iCLIP	23374342	ICLIP_hnRNPC_Hela_iCLIP_all_clusters.bedGraph.gz
hnRNPC_hg19	hnRNPC	HeLa	iCLIP	24184352	ICLIP_HNRNPC_hg19.bedGraph.gz
hnRNPL_Hela	hnRNPL	HeLa	iCLIP	24526010	ICLIP_hnRNPL_Hela_group_3975_all-hnRNPL-Hela-hg19_sum_G_hg19–ensembl59_from_2337-2339-741_bedGraph-cDNAhits-in-genome.bedGraph.gz
hnRNPL_U266	hnRNPL	U266	iCLIP	24526010	ICLIP_hnRNPL_U266_group_3986_all-hnRNPL-U266-hg19_sum_G_hg19–ensembl59_from_2485_bedGraph-cDNA-hits-in-genome.bedGraph.gz
hnRNPlike_U266	hnRNPL-like	U266	iCLIP	24526010	ICLIP_hnRNPlike_U266_group_4000_all-hnRNPLlike-U266-hg19_sum_G_hg19–ensembl59_from_2342-2486_bedGraph-cDNA-hits-in-genome.bedGraph.gz
NSUN2_293	Nsun2	HEK293	iCLIP	23871666	ICLIP_NSUN2_293_group_4007_all-NSUN2-293-hg19_sum_G_hg19–ensembl59_from_3137-3202_bedGraph-cDNA-hits-ingenome.bedGraph.gz
TDP43_hg19	TDB-43	HeLa	iCLIP	21358640	ICLIP_TDP43_hg19.bedGraph.gz
TIA1_hg19	TIA1	HeLa	iCLIP	21048981	ICLIP_TIA1_hg19.bedGraph.gz
TIAL1_hg19	TIAL1	HeLa	iCLIP	21048981	ICLIP_TIAL1_hg19.bedGraph.gz
U2AF65_Hela	U2AF2	HeLa	iCLIP	23374342	ICLIP_U2AF65_Hela_iCLIP_ctrl_all_clusters.bedGraph.gz
U2AF65_Hela_kd	U2AF2 (KD)	HeLa	iCLIP	23374342	ICLIP_U2AF65_Hela_iCLIP_ctrl+kd_all_clusters.bedGraph.gz

* All datasets, curated and partitioned by iONMF, are derived from publicly available data and can be downloaded here: https://github.com/mstrazar/iONMF/tree/master_full/datasets/clip.

Supplementary Table 11. Summary of mouse CLIP sequencing datasets

Dataset*	Target Protein	Source	Protocol	accession ID	Data file name
EZH2_mouse	EZH2	mESC	CLIP	['GSE49433', 'GSM1204909', 'GSM1204908', 'GSM1204907', 'GSM1204910']	mmCLIP_EZH2.gz
FUS_mouse	FUS	Brain & mESC-derived_neurona	CLIP	[E-MTAB-1223', 'ERR208899', 'ERR208898', 'GSE40653', 'GSM998872', 'GSM998874', 'GSM998873', 'ERR208901', 'GSE43308', 'GSM1060384', 'GSM1060385']	mmCLIP_FUS.gz
HNRNPR_mouse	HNRNPR	CD-1 & NSC-34	CLIP	['GSE77101', 'GSM2044160', 'GSM2044162', 'GSM2044163', 'GSM2044161', 'GSM2044166', 'GSM2044167', 'GSM2044165', 'GSM2044169', 'GSM2044164', 'GSM2044168']	mmCLIP_HNRNPR.gz
LIN28A_mouse	LIN28A	mESC	CLIP	['GSE37114', 'GSM910955', 'GSM910956', 'GSM910957']	mmCLIP_LIN28A.gz
RBFOX2_mouse	RBFOX2	Brain& V6.5	CLIP	['SRP128054', 'SRX3532611', 'GSE54794', 'GSM1324105', 'GSM1324104', 'SRX3532612']	mmCLIP_RBFOX2.gz
RBM10_mouse	RBM10	Mandibular & Mouse_mandibular_MEP	CLIP	['GSE89270', 'GSM2363437', 'GSM2363436', 'GSM2363438', 'GSM2363435']	mmCLIP_RBM10.gz
SRSF2_mouse	SRSF2	Fibroblast	CLIP	['GSE44591', 'GSM1088391']	mmCLIP_SRSF2.gz
SRSF3_mouse	SRSF3	P19& Embryonal	CLIP	['GSE79792', 'GSM2102849', 'GSM2102855', 'E-MTAB-747', 'ERR039836', 'GSM2102848', 'GSM2102850', 'GSM2102854', 'ERR039837', 'GSM2102847', 'GSM2102858', 'GSM2102852', 'GSM2102853', 'GSM2102856', 'GSM2102857', 'GSM2102851', 'ERR039838']	mmCLIP_SRSF3.gz
TARDBP_mouse	TARDBP	Brain	CLIP	['E-MTAB-1223', 'ERR208900', 'GSE40653', 'GSM998871', 'ERR208896', 'GSE27394', 'GSM672063', 'GSM672062', 'ERR208895']	mmCLIP_TARDBP.gz
U2AF2_mouse	U2AF2	Brain& N2A& SRRM4_KD& N2A	CLIP	['E-MTAB-1223', 'ERR208893', 'ERR208897', 'GSE57278', 'GSM1378379', 'GSM1378378']	mmCLIP_U2AF2.gz
YTHDC2_mouse	YTHDC2	Testis	CLIP	['GSE98085', 'GSM2586903', 'GSM2586904']	mmCLIP_YTHDC2.gz

* All positive sequences, curated and partitioned by CLIPdb[24], are derived from publicly available data and can be downloaded here: http://clipdb.ncrnalab.org.

Supplementary Reference

1. Zhang, Y. et al. Single-step discovery of high-affinity RNA ligands by UltraSelex. *Nat Chem Biol* (2025).
2. Zhang, H., Zhang, L., Mathews, D.H. & Huang, L. LinearPartition: linear-time approximation of RNA folding partition function and base-pairing probabilities. *Bioinformatics* **36**, i258-i267 (2020).
3. Jolma, A. et al. Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences. *Genome Res* **30**, 962-973 (2020).
4. Ameta, S., Winz, M.L., Previti, C. & Jaschke, A. Next-generation sequencing reveals how RNA catalysts evolve from random space. *Nucleic Acids Res* **42**, 1303-1310 (2014).
5. Pei, S., Slinger, B.L. & Meyer, M.M. Recognizing RNA structural motifs in HT-SELEX data for ribosomal protein S15. *BMC Bioinformatics* **18**, 298 (2017).
6. Boussebayle, A. et al. Next-level riboswitch development-implementation of Capture-SELEX facilitates identification of a new synthetic riboswitch. *Nucleic Acids Res* **47**, 4883-4895 (2019).
7. Zhang, J., Wang, L., Jaschke, A. & Sunbul, M. A Color-Shifting Near-Infrared Fluorescent Aptamer-Fluorophore Module for Live-Cell RNA Imaging. *Angew Chem Int Ed Engl* **60**, 21441-21448 (2021).
8. Whatley, A.S. et al. Potent Inhibition of HIV-1 Reverse Transcriptase and Replication by Nonpseudoknot, "UCAA-motif" RNA Aptamers. *Mol Ther Nucleic Acids* **2**, e71 (2013).
9. Camorani, S. et al. Novel Aptamers Selected on Living Cells for Specific Recognition of Triple-Negative Breast Cancer. *iScience* **23**, 100979 (2020).
10. Nguyen Quang, N., Bouvier, C., Henriques, A., Lelandais, B. & Duconge, F. Time-lapse imaging of molecular evolution by high-throughput sequencing. *Nucleic Acids Res* **46**, 7480-7494 (2018).
11. De La Fuente, A. et al. Aptamers against mouse and human tumor-infiltrating myeloid cells as reagents for targeted chemotherapy. *Sci Transl Med* **12** (2020).
12. Van Simaeys, D. et al. RNA aptamers specific for transmembrane p24 trafficking protein 6 and Clusterin for the targeted delivery of imaging reagents and RNA therapeutics to human beta cells. *Nat Commun* **13**, 1815 (2022).
13. Koyama, T., Platt, D. & Parida, L. Variant analysis of SARS-CoV-2 genomes. *Bull World Health Organ* **98**, 495-504 (2020).
14. van Zundert, G.C.P. et al. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol* **428**, 720-725 (2016).
15. Gao, Y. et al. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* **368**, 779-782 (2020).
16. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* (2024).

17. Strazar, M., Zitnik, M., Zupan, B., Ule, J. & Curk, T. Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. *Bioinformatics* **32**, 1527-1535 (2016).
18. Zarnack, K. et al. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* **152**, 453-466 (2013).
19. Konig, J. et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**, 909-915 (2010).
20. Rossbach, O. et al. Crosslinking-immunoprecipitation (iCLIP) analysis reveals global regulatory roles of hnRNP L. *RNA Biol* **11**, 146-155 (2014).
21. Hussain, S. et al. NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. *Cell Rep* **4**, 255-261 (2013).
22. Tollervey, J.R. et al. Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat Neurosci* **14**, 452-458 (2011).
23. Wang, Z. et al. iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS Biol* **8**, e1000530 (2010).
24. Yang, Y.C. et al. CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics* **16**, 51 (2015).
25. Tang, Y. et al. m6A-Atlas: a comprehensive knowledgebase for unraveling the N6-methyladenosine (m6A) epitranscriptome. *Nucleic Acids Res* **49**, D134-D143 (2021).
26. Yang, Y. et al. Deciphering 3'UTR Mediated Gene Regulation Using Interpretable Deep Representation Learning. *Adv Sci (Weinh)*, e2407013 (2024).
27. Wang, Z. et al. Deciphering cell lineage specification of human lung adenocarcinoma with single-cell RNA sequencing. *Nat Commun* **12**, 6500 (2021).
28. Kaiser, A.M. et al. p53 governs an AT1 differentiation programme in lung cancer suppression. *Nature* **619**, 851-859 (2023).